**Car price prediction project**

***I would like to express my special thanks of gratitude to my organization (Fliprobo))who gave me the golden opportunity to do this wonderful project on the topic (*Car price prediction project *), which also helped me in doing a lot of Research and i came to know about so many new things I am really thankful to them.***

About the problem:

With the covid 19 impact in the market, we have seen a lot of changes in the car market. Now some

cars are in demand hence making them costly and some are not in demand hence cheaper. One

Our clients work with small traders, who sell used cars. With the change in market due to covid

19 impact, our client is facing problems with their previous car price valuation machine learning

models. So, they are looking for new machine learning models from new data. We have to make

car price valuation model.

**Solution to above problem:**

Importing selenium library and performing web scraping from the pre owned car’s website

## What is Web Scraping? :-

Web Scrapping also called “Crawling” or “Spidering” is the technique to gather data automatically from an online source usually from a website. While Web Scrapping is an easy way to get a large volume of data in a relatively short time frame, it adds stress to the server where the source is hosted.

## How Web Scraping is useful? –

Web scraping can help us extract an enormous amount of data about customers, products, people, stock markets, etc.

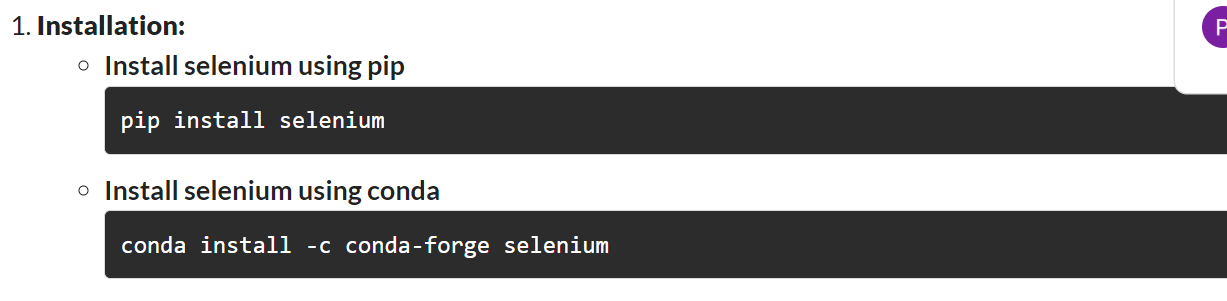
One can utilize the data collected from a website such as e-commerce portal, Job portals, social media channels to understand customer’s buying patterns, employee attrition behavior, and customer’s sentiments and the list goes on.

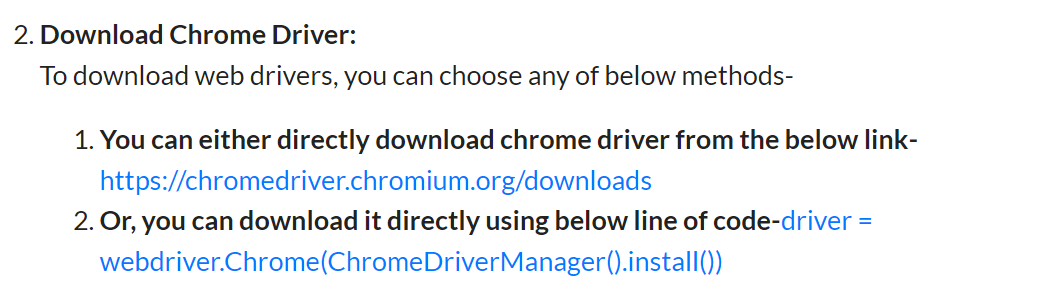
Most popular libraries or frameworks that are used in Python for Web – Scrapping are BeautifulSoup, Scrappy & Selenium.

## What is Selenium: –

Selenium is an open-source web-based automation tool. Selenium primarily used for testing in the industry but It can also be used for web scraping. We’ll use the Chrome browser but you can try on any browser, It’s almost the same.

#### Setup & tools:-





Following methods will help us to find elements in a Web-page (these methods will return a list):

* find\_elements\_by\_name
* find\_elements\_by\_xpath
* find\_elements\_by\_link\_text
* find\_elements\_by\_partial\_link\_text
* find\_elements\_by\_tag\_name
* find\_elements\_by\_class\_name
* find\_elements\_by\_css\_selector



Scraping the data for car names



Scraping the data for price



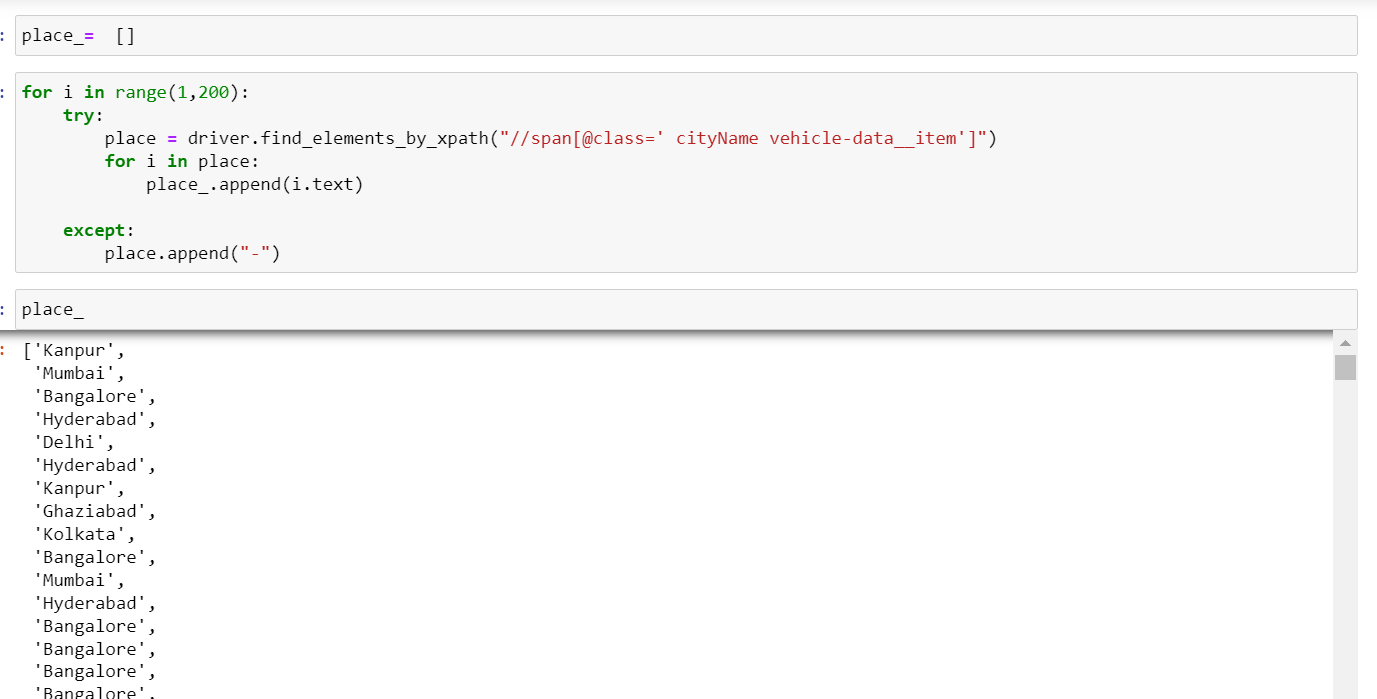
Scraping the range of km that cars are driven



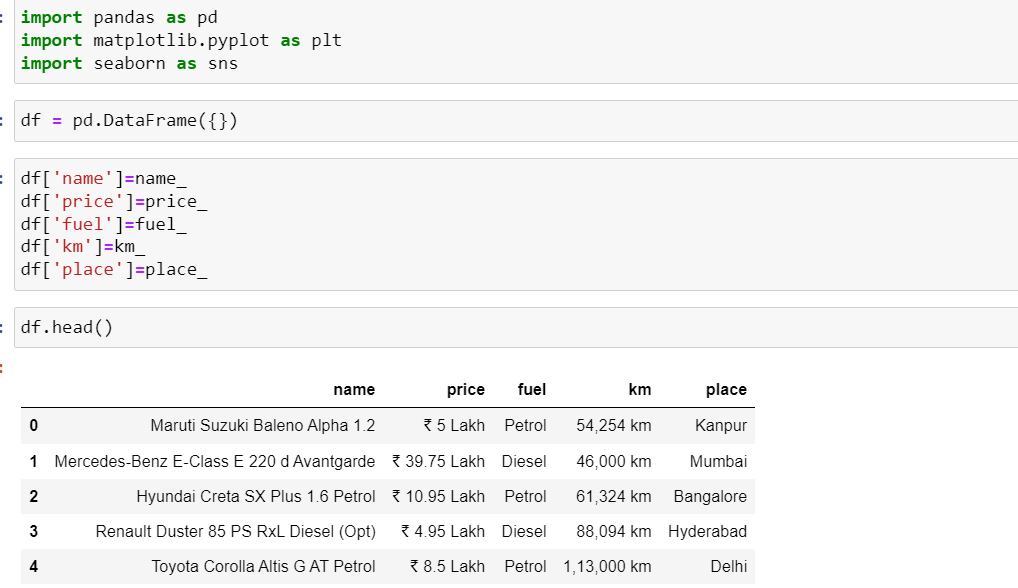
Scraping the type of fuel cars are available



Scarping the Car’s location



Now importing pandas library for making the data frame of the above scraped data and also importing some visualisation tools and library like matplotlib and seaborn

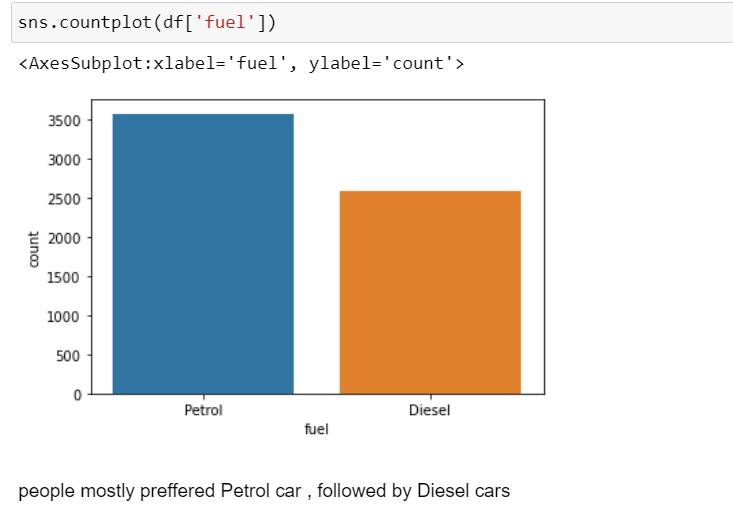


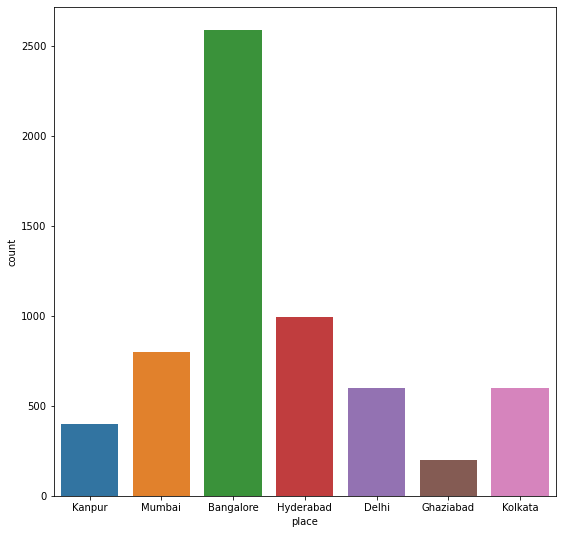
Now performing some EDA (exploratory data analysis ) to gain some insights from the scraped data

About EDA

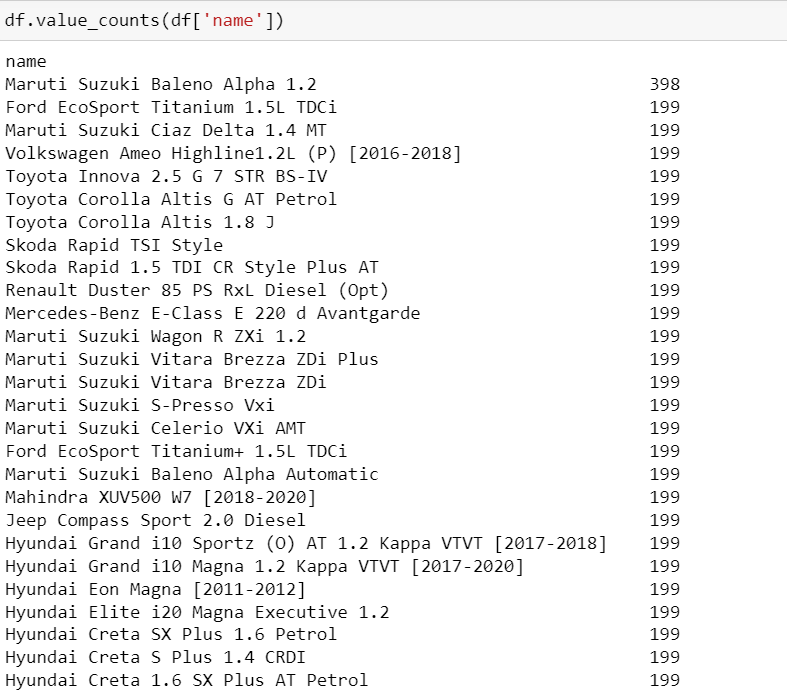
Exploratory Data Analysis, or EDA, is an important step in any Data Analysis or Data Science project. EDA is the process of investigating the dataset to discover patterns, and anomalies (outliers), and form hypotheses based on our understanding of the dataset.

EDA involves generating summary statistics for numerical data in the dataset and creating various graphical representations to understand the data better. In this article, we will understand EDA with the help of an example dataset. We will use Python language (Pandas library) for this purpose.

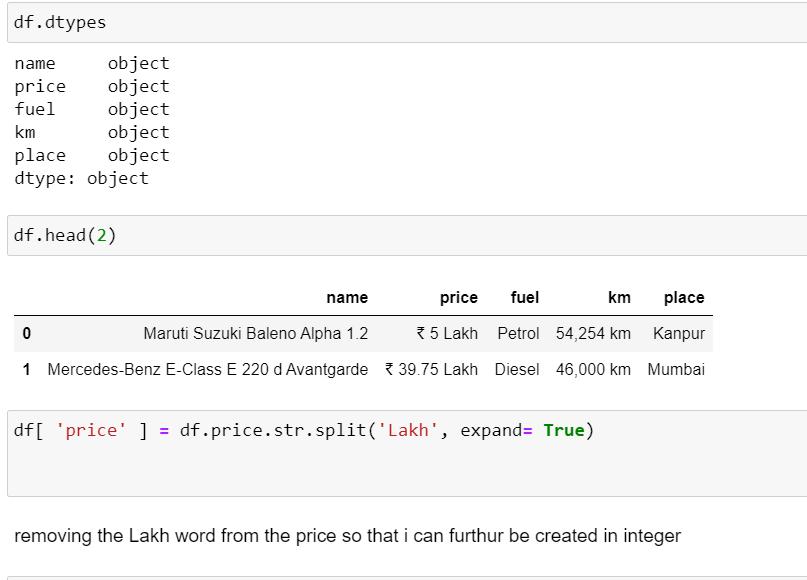


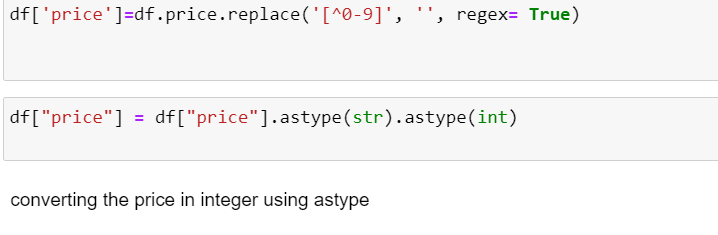


As per the collected data mostly used cars are traded in Bangalore , Mumbai , Hyderabad , delhi and least are traded in Kanpur , Ghaziabad , kolkata



These are the above cars that are scrapped from the website and also i which amount particular car is scrapped







Above is the statistical information about the data and they are as follows defined below :

max count of car name is = 6369

mean = 14.45

standard deviation = 8.5

minimum value = 0

max values = 29

max count of Price is = 6369

mean = 646.25

standard deviation = 751.12

minimum value = 0

max values = 3975

max count of car fuel is = 6369

mean = 0.41

standard deviation = 0.49

minimum value = 0

max values = 1

max count of km is = 6369

mean = 46003

standard deviation = 24184

minimum value = 0

max values = 128000.0

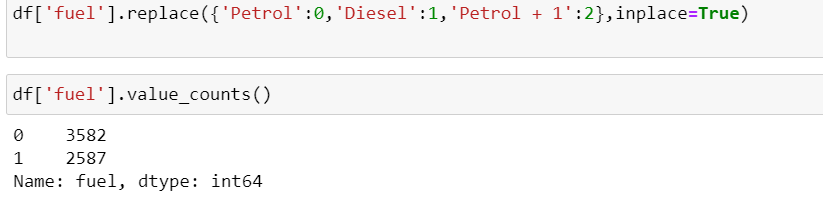
max count of place is = 6368

mean =3.0

standard deviation = 1.7

minimum value = 0.0

max values = 7.0

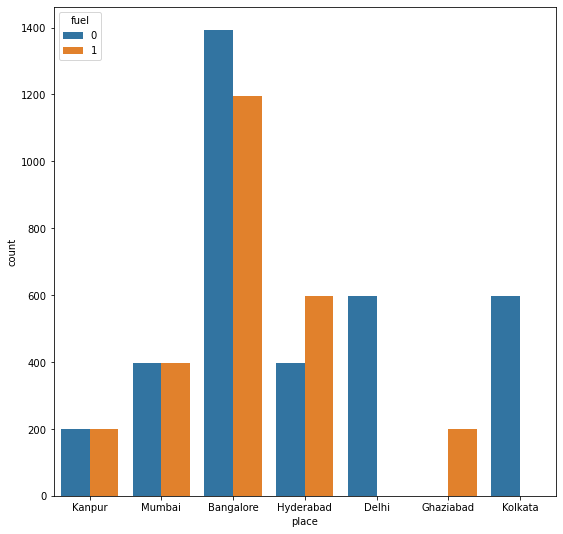


replacing the values of

petrol : 0

Diesel :1

LPG : 3

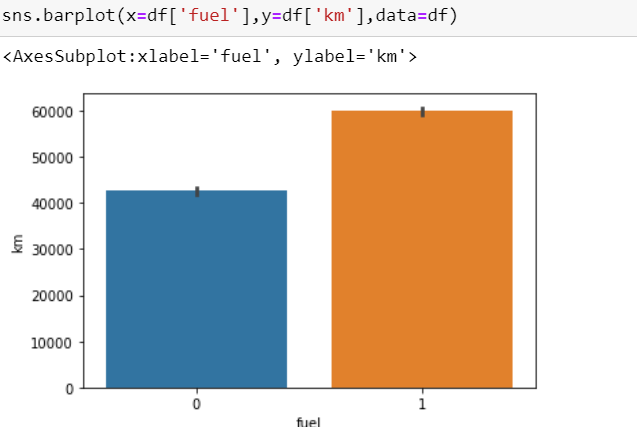


as per the data

petrol cars are more in demand as compared to diesel cars also the demand of LPG cars is the least

Banglore shows the petrol demand highest in number followed by delhi , kolkata , Hyderabad , mumbai , kanpur

also about diesel the demand is high in Bangalore , Hyderabad , mumbai , kanpur , Ghaziabad



usually in pre owned cars

the petrol cars are less driven as compared to diesel cars

As per the data Diesel cars are more costly or has more price as compared to petrol cars

if we see the places with high cost they are as follows arranged in descending order

1.Mumbai

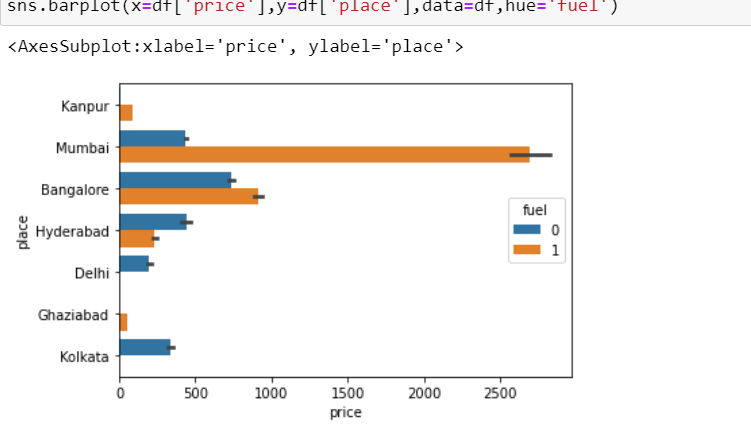
2.Banglore

3.Kolkata

4.Hyderabad

5.Delhi

6.Kanpur



mostly diesel cars are high in cost as follows

1.Mumbai

2.Banglore

3.Kanpur

4.Ghaziabad

And if we talk about petrol cars cost as per the places they are as follows

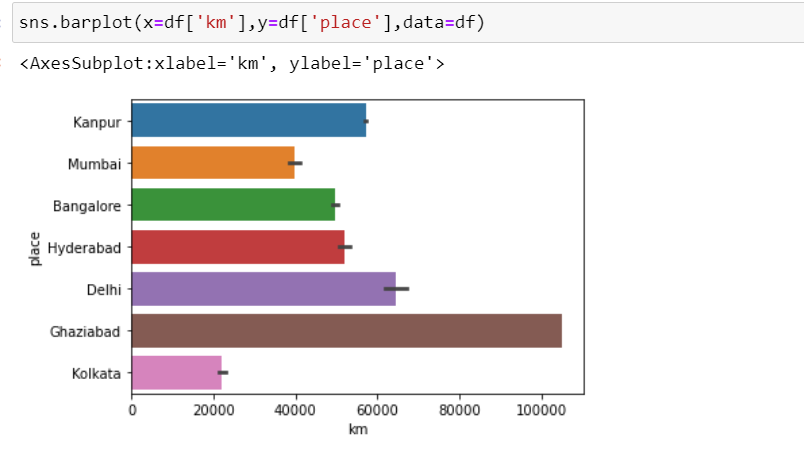
1.Banglore

2.Mumbai

3.Hyderabad

4.Delhi

5.Kolkata



this bar plot shows the max or min car driver at various place and these are as follows arranged in descending order

1.Ghaziabad

2.Delhi

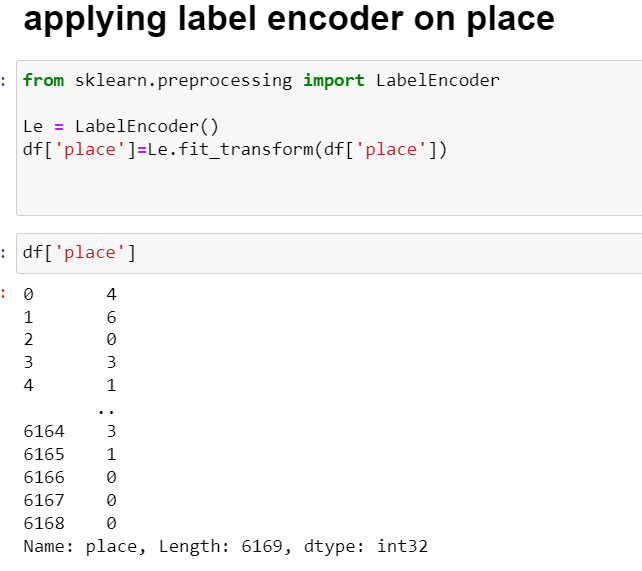
3.Hyderabad

4.Kanpur

5.Bangalore

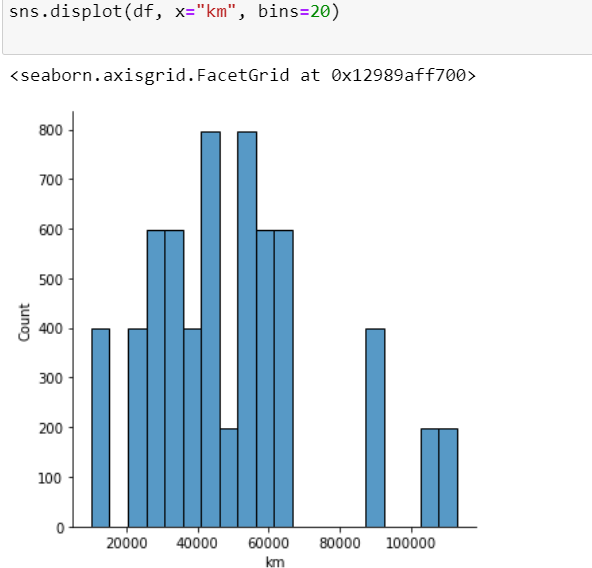
6.Mumbai

7.Kolkata



**What is label encoder**

Label Encoding refers to converting the labels into a numeric form so as to convert them into the machine-readable form. Machine learning algorithms can then decide in a better way how those labels must be operated. It is an important preprocessing step for the structured dataset in supervised learning.

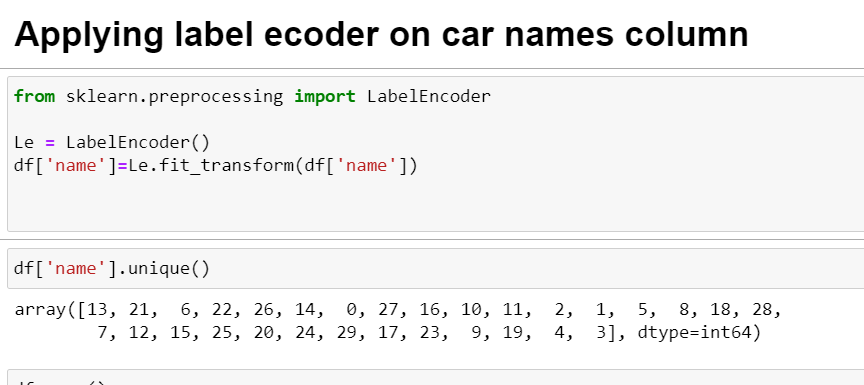


as per the figure

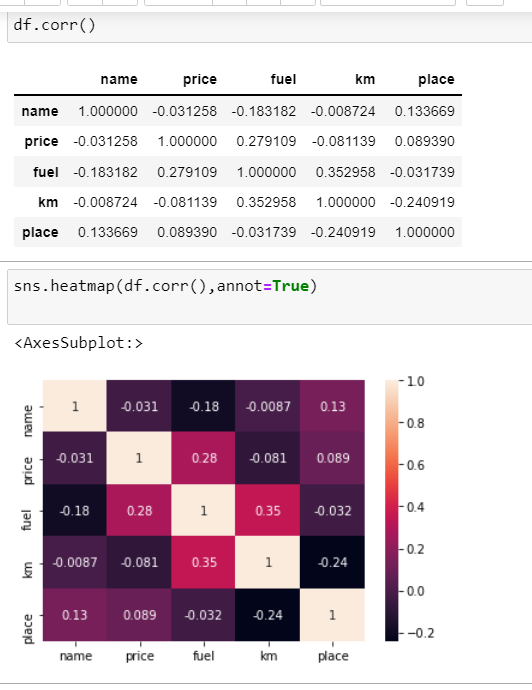
max cars are driven between 40,000 km to 60,000 km

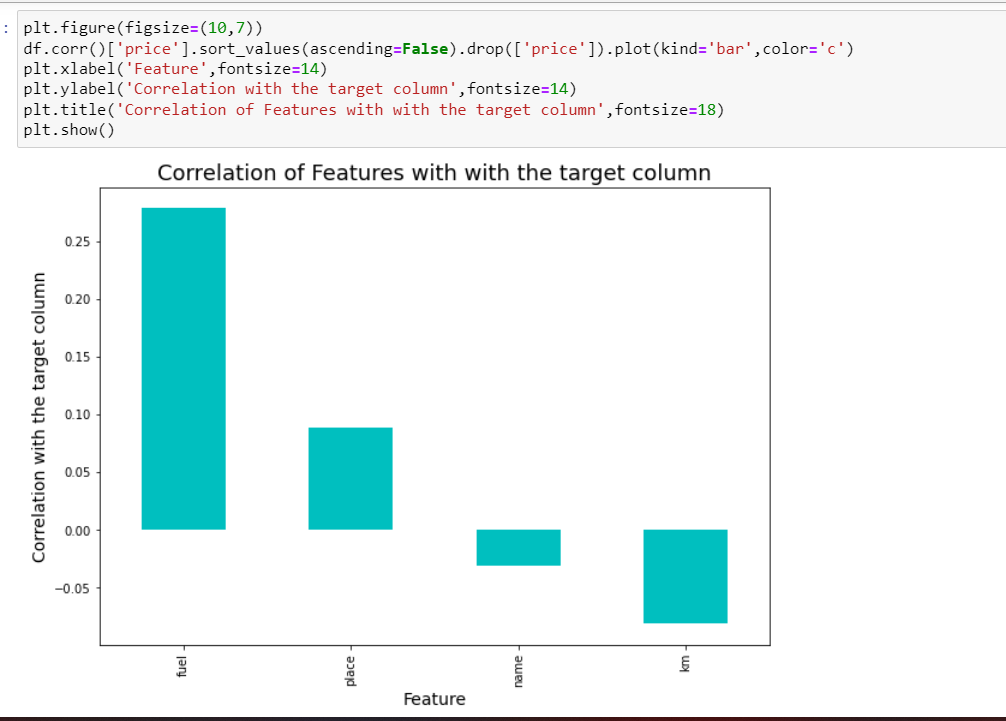
min cars driven around 10,000 km to 20,000 km

max car driven is more than 1,00,000 km



**Checking the correlation in relation to price:**





This figure shows the correlation and as per given data

max positive correlation is shown between price and fuel then

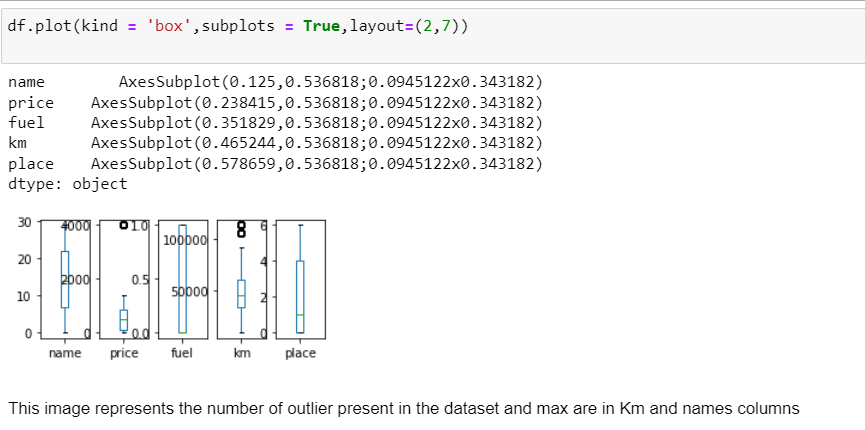
positive correlation is shown between price and place then

positive correlation is shown between price and name then

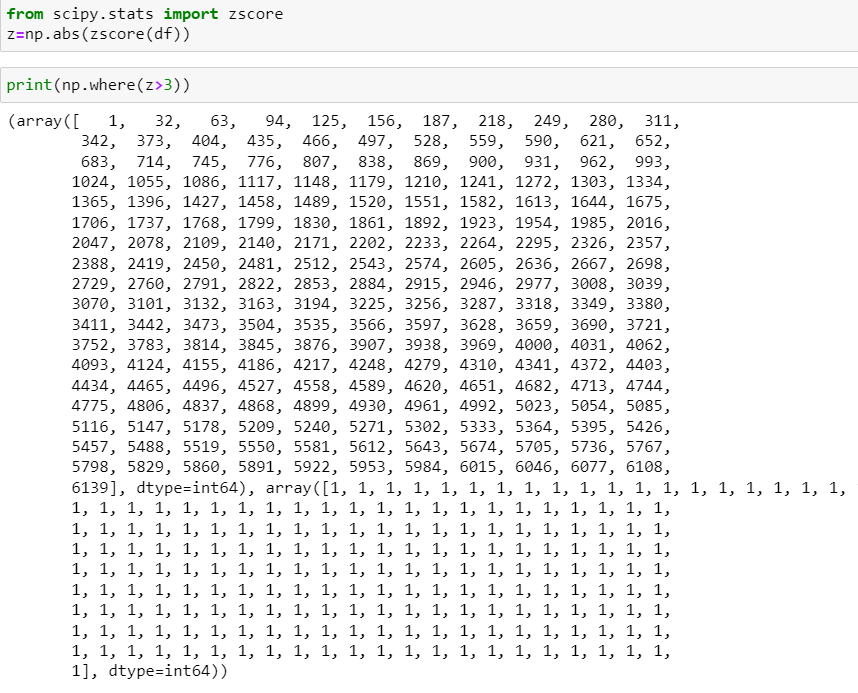
and at last there is negative correlation shows between km driven and price

***About outliers***

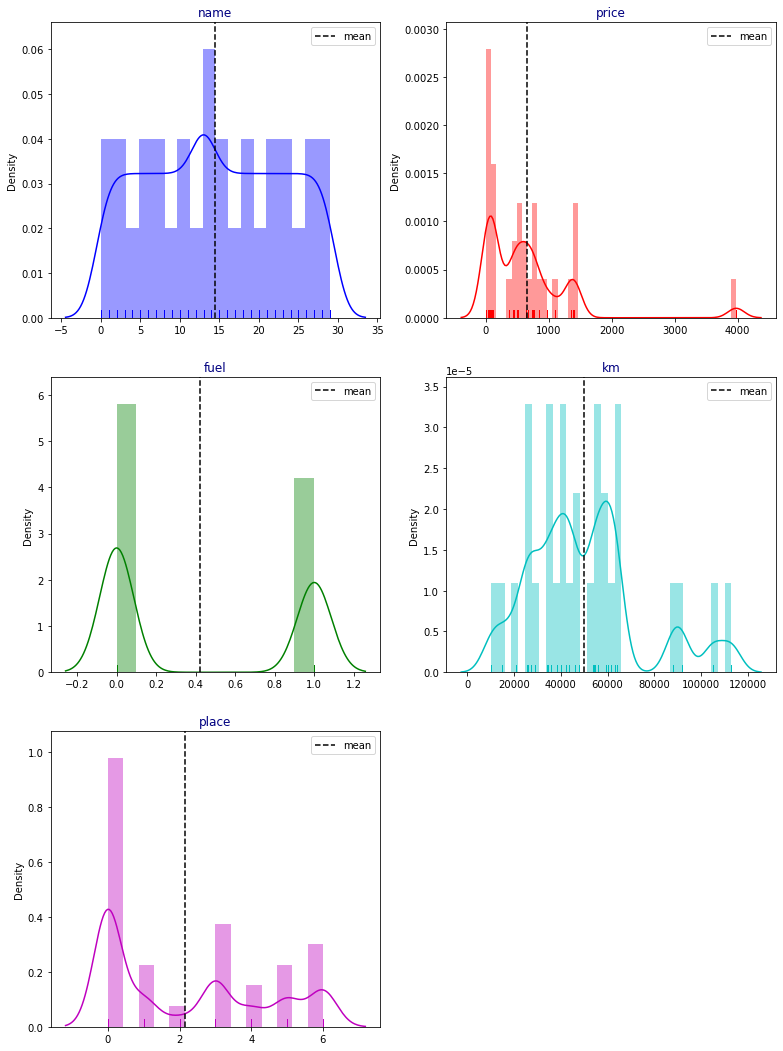
An outlier is an object that deviates significantly from the rest of the objects. They can be caused by measurement or execution error. The analysis of outlier data is referred to as outlier analysis or outlier mining.



***Removing the outlier using library Zscore***

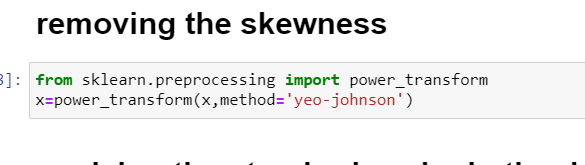
***Around 7 % of the outlier were present in the data set and were removed*** 

***Then checking the skewness from the data***



About skewness

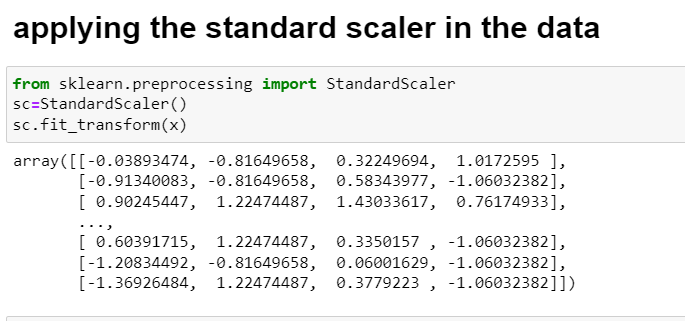
Skewness, in statistics, is the degree of asymmetry observed in a probability distribution. Distributions can exhibit right (positive) skewness or left (negative) skewness to varying degrees. A normal distribution (bell curve) exhibits zero skewness.



***Scaling the data***

**About standard scaler**

Data Scaling is a data preprocessing step for numerical features. Many machine learning algorithms like Gradient descent methods, KNN algorithm, linear and logistic regression, etc. require data scaling to produce good results. Various scalers are defined for this purpose. This article concentrates on Standard Scaler and Min-Max scaler. The task here is to discuss what they mean and how they are implemented using in-built functions that come with this package



***Now applying various algorithms for predicting the price of cars***

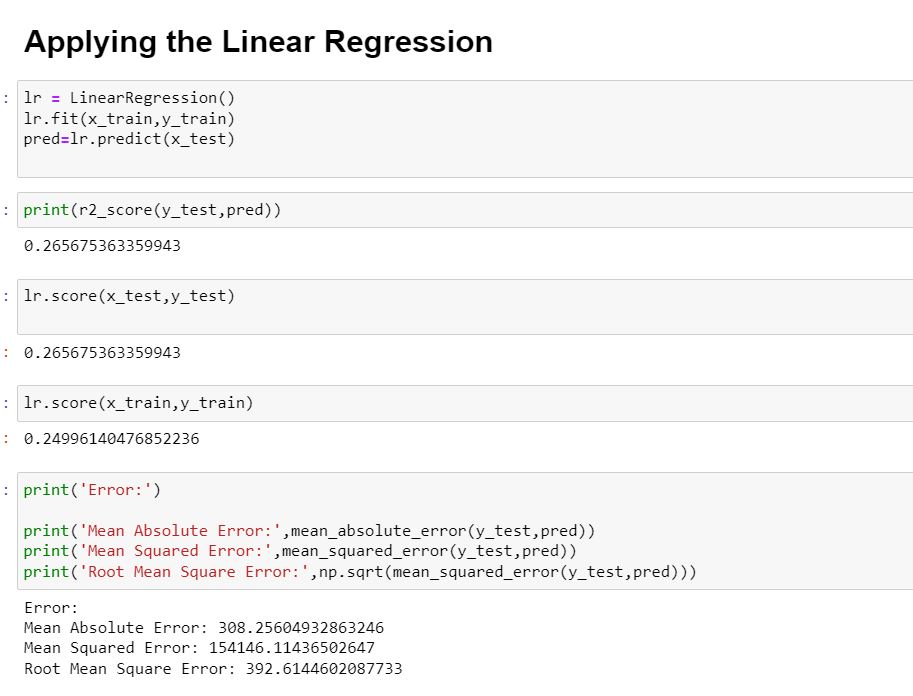
**1 linear regression**

In the most simple words, Linear Regression is the supervised Machine Learning model in which the model finds the best fit linear line between the independent and dependent variable i.e it finds the linear relationship between the dependent and independent variable.

Linear Regression is of two types: Simple and Multiple. Simple Linear Regression is where only one independent variable is present and the model has to find the linear relationship of it with the dependent variable

Whereas, In Multiple Linear Regression there are more than one independent variables for the model to find the relationship.

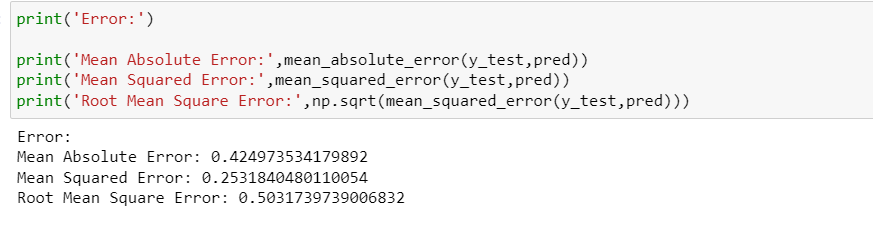
Equation of Simple Linear Regression, where bo is the intercept, b1 is coefficient or slope, x is the independent variable and y is the dependent variable.



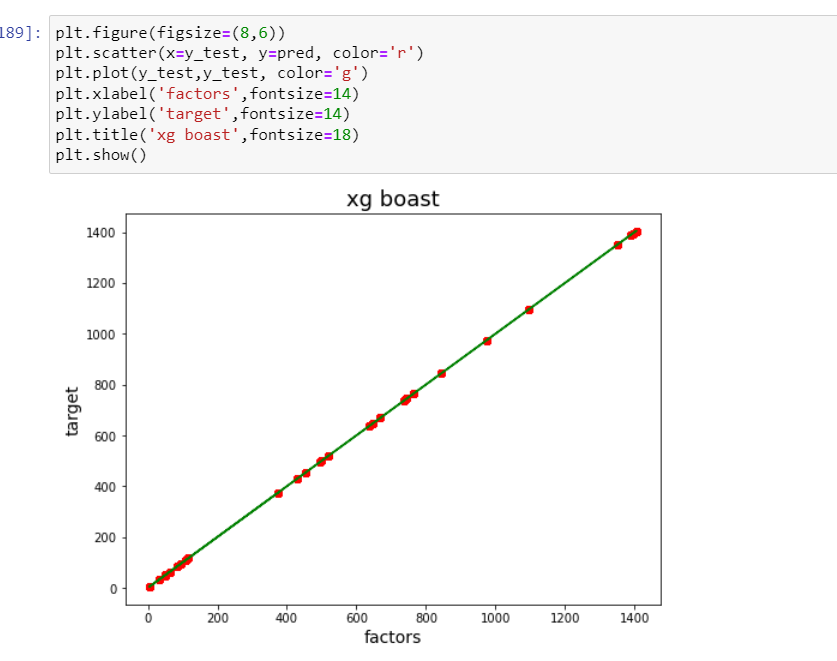
**Then applying XG Boost**

When using [gradient boosting for regression](https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost-HowItWorks.html), where the weak learners are considered to be regression trees, each of the regression trees maps an input data point to one of its leaves that includes a continuous score. XGB minimises a regularised objective function that merges a convex loss function, which is based on the variation between the target outputs and the predicted outputs. The training then proceeds iteratively, adding new trees with the capability to predict the residuals as well as errors of prior trees that are then coupled with the previous trees to make the final prediction.





**Plotting of the XG Boost**



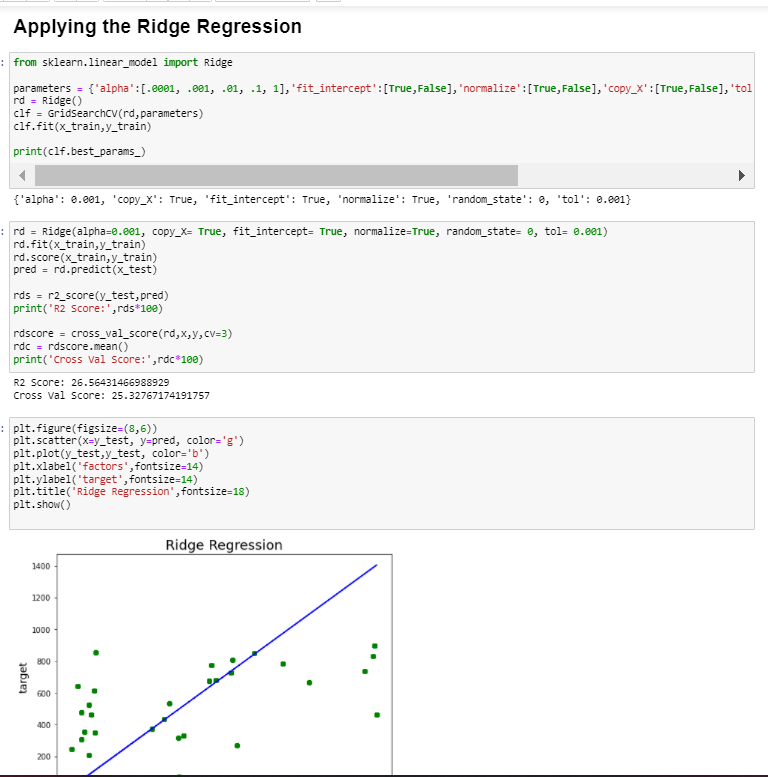
***Now applying ridge and lasso regression***

**Ridge Regression :**

In Ridge regression, we add a penalty term which is equal to the square of the coefficient. The *L2* term is equal to the square of the magnitude of the coefficients. We also add a coefficient to control that penalty term. In this case if is zero then the equation is the basic OLS else if then it will add a constraint to the coefficient. As we increase the value of this constraint causes the value of the coefficient to tend towards zero. This leads to both low variance (as some coefficient leads to negligible effect on prediction) and low bias (minimization of coefficient reduces the dependency of prediction on a particular variable).

**Lasso Regression :**

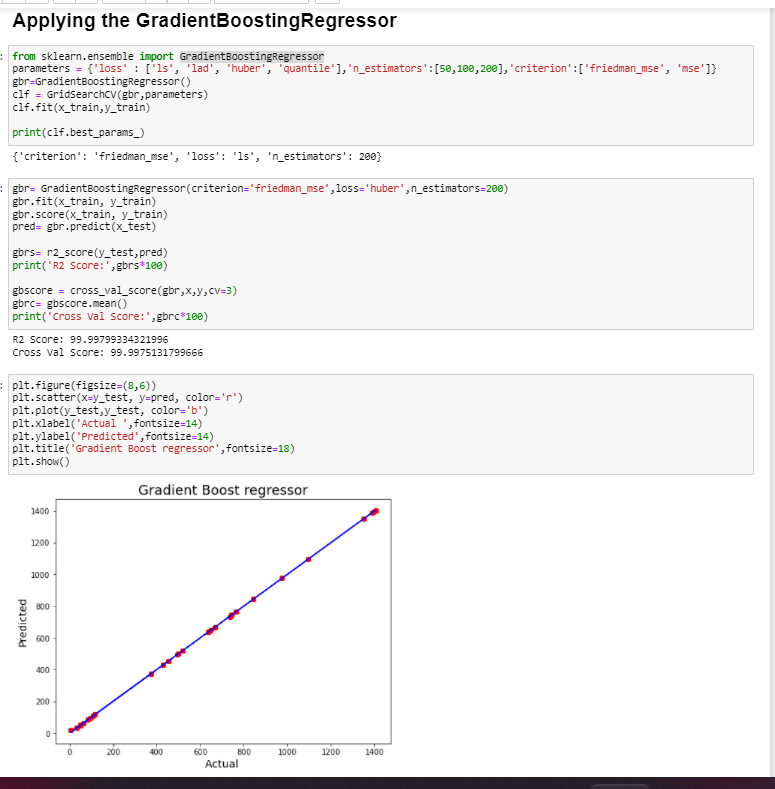
Lasso regression stands for Least Absolute Shrinkage and Selection Operator. It adds a penalty term to the cost function. This term is the absolute sum of the coefficients. As the value of coefficients increases from *0* this term penalizes, causing the model to decrease the value of coefficients in order to reduce loss. The difference between ridge and lasso regression is that it tends to make coefficients to absolute zero as compared to Ridge which never sets the value of coefficient to absolute zero.





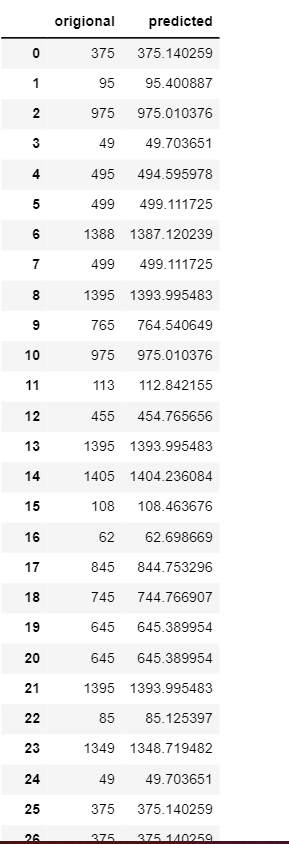
**About gradient boosting**

Gradient Boosting is a popular boosting algorithm. In gradient boosting, each predictor corrects its predecessor’s error. In contrast to Adaboost, the weights of the training instances are not tweaked, instead, each predictor is trained using the residual errors of the predecessor as labels.



hence xgboost is performing the best so saving and using xg boost for making future predictions

**Example of our predicted results**



***Observation and conclusion***

* There is a huge demand of pre owned cars
* Now people usually prefer petrol car over diesel car
* Petrol car proved to be more cheaper in price as compared to diesel cars
* There is a great demand of both Diesel and petrol cars in south india as compared to north
* North indian car dealer may put their major focus over the petrol cars
* There are some cities which can proved to be profitable if sales are increased
* Maximum cars are demand between 30,000 km to 50,000 km